Neuroscience and Biobehavioral Reviews 74 (2017) 58–75

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications

Sandra Vieira[a,*], Walter H.L. Pinaya[b], Andrea Mechelli[a]

[a] Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park, SE5 8AF, United Kingdom
[b] Centre of Mathematics, Computation, and Cognition, Universidade Federal do ABC, Rua Arcturus, Jardim Antares, São Bernardo do Campo, SP CEP 09.606-070, Brazil

ARTICLE INFO

Article history:
Received 2 October 2016
Received in revised form
22 December 2016
Accepted 4 January 2017
Available online 10 January 2017

Keywords:
Deep learning
Machine learning
Neuroimaging
Pattern recognition
Multilayer perceptron
Autoencoders
Convolutional neural networks
Deep belief networks
Psychiatric disorders
Neurologic disorders

ABSTRACT

Deep learning (DL) is a family of machine learning methods that has gained considerable attention in the scientific community, breaking benchmark records in areas such as speech and visual recognition. DL differs from conventional machine learning methods by virtue of its ability to learn the optimal representation from the raw data through consecutive nonlinear transformations, achieving increasingly higher levels of abstraction and complexity. Given its ability to detect abstract and complex patterns, DL has been applied in neuroimaging studies of psychiatric and neurological disorders, which are characterised by subtle and diffuse alterations. Here we introduce the underlying concepts of DL and review studies that have used this approach to classify brain-based disorders. The results of these studies indicate that DL could be a powerful tool in the current search for biomarkers of psychiatric and neurologic disease. We conclude our review by discussing the main promises and challenges of using DL to elucidate brain-based disorders, as well as possible directions for future research.

Contents

* Corresponding author.
E-mail address: sandra.vieira@kcl.ac.uk (S. Vieira).

http://dx.doi.org/10.1016/j.neubiorev.2017.01.002

1. Introduction

In the last two decades, neuroimaging studies of psychiatric and neurological patients have relied on mass-univariate analytical techniques (e.g. statistical parametric mapping). These studies typically compared patients with a diagnosis of interest against disease-free individuals and reported neuroanatomical or neurofunctional differences at group level. The simplicity and interpretability of this approach have led to significant advances in our understanding of the neurobiology of psychiatric and neurological disorders. Mass-univariate analytical techniques, however, suffer from at least two significant limitations. First, statistical inferences are drawn from multiple independent comparisons (i.e. one for each voxel) based on the assumption that different brain regions act independently. This assumption, however, is not in line with our current understanding of brain function in health and disease (Fox et al., 2005; Biswal et al., 2010); for example, several psychiatric and neurological symptoms are best explained by network-level changes in structure and function rather than focal alternations (Mulders et al., 2015; Kennedy and Courchesne, 2008; Sheffield and Barch, 2016). Second, mass-univariate techniques can be used to detect differences between groups but do not allow statistical inferences at the level of the individual. In contrast, a clinician has to make diagnostic and treatment decisions about the person in front of them. These two limitations may have contributed to the limited translational impact of neuroimaging findings in everyday clinical practice so far.

In an attempt to overcome these limitations, the neuroimaging community has developed a growing interest in machine learning (ML), an area of artificial intelligence that aims to develop algorithms that discover trends and patterns in existing data and use this information to make predictions on new data. This is achieved through the use of computational statistics and mathematical optimization (Hastie et al., 2001). ML methods are multivariate and therefore take the inter-correlation between voxels into account, thereby overcoming the first limitation of mass-univariate analytical techniques. In addition, ML methods allow statistical inferences at single subject level and therefore could be used to inform diagnostic and prognostic decisions of individual patients, thereby overcoming the second limitation of mass-univariate analytical techniques (Arbabshirani et al., 2016). ML methods can be divided into two broad categories: supervised and unsupervised learning. In supervised ML, one seeks to develop a function which maps two or more sets of observations to predefined categories or values. In contrast, unsupervised methods seek to determine how the data are organized without using any a priori information supplied by the operator; here the main objective is to discover unknown structure in the data (Hastie et al., 2001).

Over the past decade, several ML methods have been applied to neuroimaging data from psychiatric and neurological patients with varying degrees of success (Arbabshirani et al., 2016; Wolfers et al., 2015). The most popular amongst these methods is Support Vector Machine (SVM), a supervised technique that works by estimating an optimal hyperplane that best separates two classes. When these classes are not linearly separable, SVM uses external functions (kernels) that map the original data into a new feature space where the data become linearly separable (Pereira et al., 2009; Vapnik, 1995). Despite its popularity, SVM has been criticised for not performing well on raw data and requiring the expert use of design techniques to extract the less redundant and more informative features (a step known as "feature selection") (LeCun et al., 2015; Plis et al., 2014). These features, rather than the original data, are then used for classification. While SVM remains a very popular technique within the neuroimaging community, an alternative family of ML methods known as deep learning (DL) (Bengio, 2009) is gaining considerable attention in the wider scientific community (Arbabshirani et al., 2016; Calhoun and Sui, 2016; LeCun et al., 2015). Deep learning methods are a type of representation-learning methods, which means that they can automatically identify the optimal representation from the raw data without requiring prior feature selection. This is achieved through the use of a hierarchical structure with different levels of complexity, which involves the application of consecutive nonlinear transformations to the raw data. These transformations result in increasingly higher levels of abstraction, where higher-level features are more invariant to the noise present in the input data than lower level ones (LeCun et al., 2015). Inspired by how the human brain processes information, the building blocks of DL neural networks – known as "artificial neurons" – are loosely modelled after biological neurons. Artificial neurons are organized in layers. A deep neural network consists of an input layer, two or more hidden layers and an output layer. The input layer comprises the data inputted into the model (e.g. voxel intensity); the hidden layers learn and store increasingly more abstract features of the data; these features are then fed to the output layer that assigns the observations to classes (e.g. controls vs. patients). Learning is achieved through an iterative process of adjustment of the interconnections between the artificial neurons within the network, much like in the human brain (Bengio, 2009). An essential aspect of DL that differentiates it from other ML methods is that the features are not manually engineered; instead, they are learned from the data, resulting in a more objective and less bias-prone process. Besides, the ability to achieve higher orders of abstraction and complexity relative to other ML methods such as SVM makes DL better suited for detecting complex, scattered and subtle patterns in the data (Plis et al., 2014).

From a historical perspective, the use of DL in scientific research can be traced back to the perceptron (i.e. the original version of the artificial neuron), which many researchers refer to as the first ML algorithm (McCulloch and Pitts, 1943). After several setbacks, the pioneering work of Warren McCulloch and Walter Pitts resulted in the development of what is now known as artificial neural networks. However, such networks were able to handle a limited number of hidden layers. It was only in the 2000s that researchers developed a new approach for training artificial neural networks that allowed the inclusion of several hidden layers resulting in greater levels of complexity (Hinton et al., 2006). This breakthrough led to the development of a new family of ML methods – known as deep learning – which has been shown to outperform previous state-of-the-art classification methods in areas such as speech recognition, computer vision and natural language processing (Krizhevsky et al., 2012; Le et al., 2012).

The use of DL could be particularly useful in the investigation of psychiatric and neurological disorders, which tend to be associated with subtle and diffuse neuroanatomical and neurofunctional abnormalities. Since high-level features can be more robust against noise in the input data, deep architectures may be more suitable to identify diagnostic and prognostic biomarkers than conventional ML methods. DL techniques might also provide an ideal tool to investigate the multi-faceted nature of psychiatric and neurological disorders since cross-modality relationships (e.g. neuroimaging and genetics) are likely to occur at an even deeper level (Plis et al., 2014). In addition to these conceptual differences, the use of DL to investigate psychiatric and neurological disorders has the practical advantage of not requiring manual feature selection (LeCun et al., 2015). Therefore, it is unsurprising that an increasing number of neuroimaging studies are using DL to elucidate the neural correlates of these disorders (e.g. Payan and Montana, 2015; Plis et al., 2014; Kim et al., 2016).

Given the insurgence of interest in DL within the field of neuroimaging, this review aims to give a brief overview of DL and potential applications to the investigation of brain-based disorders. In the first part of the review, we outline the underlying concepts of DL. To achieve this, we will use one of the simplest DL structures, i.e. the multilayer perceptron, to illustrate the steps of training and testing. This will be followed by a brief description of the most common DL architectures used in the field of neuroimaging, including stacked autoencoders, deep belief networks and convolutional neural networks. The second part of this article aims to summarise the studies that have applied DL to neuroimaging data to investigate psychiatric and neurological disorders. Finally, in the third part of the review, we discuss the main themes that have emerged from our review of the existing literature, and make a number of suggestions for future research directions.

2. Overview

Deep learning refers to the training and testing of multi-layered neural networks that are capable of learning complex structures and achieve high levels of abstraction. There are two main types of DL models which differ with respect to how the information is propagated through the network. In feedforward networks, the information is propagated through the network in just one direction, from the input to the output layer. Recurrent networks, in contrast, contain feedback connections that allow the information from past inputs to affect the current output. These connections enable the information to persist within the neural network, akin to a form of memory, and this allows the models to process sequential data, such as speech and language, in a natural way.

The implementation of DL in the context of supervised classification problems involves two main steps. In the first step, the so-called *training phase*, a subset of the available data known as the *training set* is used to optimize the network's parameters to perform the desired task (classification). In the second step, the so-called *testing phase*, the remainder subset which is known as the *test set* is used to assess whether the trained model can blind-predict the class of new observations. When the amount of available data is limited, it is also possible to run the training and testing phases several times on different training and test splits of the original data and then estimate the average performance of the model – an approach known as cross-validation. The two phases of training and testing are not a specific feature of DL but are used in conventional ML methods.

In this section, we will discuss the use of feedforward DL for classification problems. We will start with the multilayer perceptron (MLP), the simplest deep neural network (DNN) architecture, to illustrate three important aspects of deep learning – network structure, training and testing. We will then describe more complex networks, including stacked autoencoders and deep belief networks. Finally, we will describe the increasingly popular convolutional neural networks (CNN), an important adaptation of the MLP that has come to be considered the state-of-the-art for computer vision.

2.1. Multilayer perceptron

2.1.1. Network structure

MLPs are organized in a layer-wise structure where each layer stores increasingly more abstract representations of the data (Fig. 1). The first layer is the input layer where the data is entered into the model. In neuroimaging, the data can be represented as a one-dimensional vector with each value corresponding to the intensity of one voxel. The last layer is the output layer which, in the context of classification, yields the probability of a given subject belonging to one group or the other. The layers between the input and output layers are called hidden layers, with the number of hidden layers representing the depth of the network. Each layer comprises a set of artificial neurons or "nodes" (Fig. 1a) in which each neuron is fully connected to all neurons in the previous layer (Fig. 1b). Each connection is associated with a weight value, which reflects the strength and direction (excitatory or inhibitory) of each neuron input, much like a synapse between two biological neurons.

Unlike SVM, which relies on expert designed transformations to handle nonlinearly separable classes, the structure of neural networks itself allows the transformation of the input space. The consecutive layers perform a cascade of nonlinear transformations that distort the input space allowing the data to become more easily separable (Fig. 2). The optimal number of layers and nodes within each layer are not estimated as part of the learning process itself but are defined *a priori*. These *a priori* parameters, which are not optimized during the training, are called hyperparameters. It should be noted that the development of algorithms to find optimum values of these hyperparameters is an active area of research, and that at present there are no fixed rules (Bergstra et al., 2011; Gelbart et al., 2014).

2.1.2. Training

Traditionally, neural networks can learn through a gradient descent-based algorithm. The gradient descent algorithm aims to find the values of the network weights that best minimise the error (difference) between the estimated and true outputs. Since MLPs can have several layers, in order to adjust all the weights along the hidden layers, it is necessary to propagate this error backward (from the output to the input layer). This propagation procedure is called backpropagation, and allows the network to estimate how much the weights from the lower layers need to be changed by the gradient descent algorithm. Initially, when a neural network is trained,

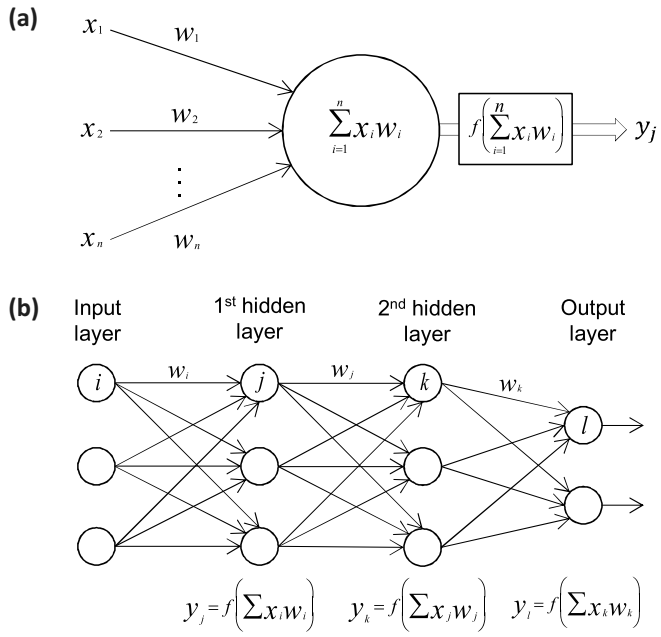

Fig. 1. (a) The building block of deep neural networks – artificial neuron or node. Each input x_i has an associated weight w_i. The sum of all weighted inputs, $\Sigma x_i w_i$, is then passed through a nonlinear activation function f, to transform the pre-activation level of the neuron to an output y_j. For simplicity, the bias terms have been omitted. The output y_j then serves as input to a node in the next layer. Several activation functions are available, which differ with respect to how they map a pre-activation level to an output value. The most commonly activation functions used are the rectifier function (where neurons that use it are called rectified linear unit (ReLU)), the hyperbolic tangent function, the sigmoid function and the softmax function. The latter is commonly used in the output layer as it can compute the probability of multiclass labels. (b) Example of a feedforward multilayer neural network (also referred to as multilayer perceptron) with two classes, in which the nodes in one layer are connected to all neurons in the next layer (fully connected network). For each neuron j in the first hidden layer, a nonlinear function is applied to the weighted sum of the inputs. The result of this transformation (y_j) serves as input for the second hidden layer. The information is propagated through the network up to the output layer, where the softmax function yields the probability of a given observation belonging to each class.

the weights are set at random. When the training set is presented to the network, this forward propagates the data through the nonlinear transformation along the layers. The estimated output is then compared to the true output, and the error is propagated from the output towards the input, allowing the gradient descent algorithm to adjust the weights as required. The process continues iteratively until the error has reached its minimum value. The backpropagation algorithm does not work well with the original models of DNNs that were based on sigmoid and hyperbolic tangent nonlinearities. In these models, the information of the error becomes increasingly smaller as it propagates backward from the output to the input layer, to a point where initial layers do not get useful feedback on how to adjust their weights – an issue known as the vanishing gradient problem. Therefore, initially, the use of backpropagation yielded poor solutions for networks with three or more hidden layers (Schmidhuber, 2015). In 2006, however, Hinton and colleagues put forward the idea of "greedy layerwise training", which consists of two steps: 1) an unsupervised step, where each layer is trained individually and 2) a supervised step, where the previously trained layers are stacked, one additional layer is added to perform the classification (the output layer), and the whole network parameters are fine-tuned (Hinton et al., 2006). This breakthrough led to the fast-growing interest in deep learning and enabled the development of at least two types of pre-trained networks that have shown promising results: stacked autoencoders and deep belief networks. It should be noted that these methods are not actual classifiers themselves; instead, they are networks that are pre-trained to learn useful patterns in the data and then fed to a real classifier at the final layer. These two types of networks and their unique characteristics are described in Section 2.2 and 2.3.

2.1.3. Testing

The performance of a deep neural network can be evaluated by several performance measures, such as sensitivity, specificity, accuracy and F-score. Sensitivity refers to the proportion of true positives correctly identified (e.g. the proportion of subjects that were predicted as patient and are true patients), and specificity refers to true negatives correctly identified (e.g. the proportion of subjects that were predicted as healthy controls and are true healthy controls). The accuracy of a classifier represents the overall proportion of correct classifications. The statistical significance of this overall accuracy can be tested using parametric tests such as permutation testing, which measures how likely the observed accuracy would be obtained by chance. Metrics such as F-score and balanced accuracy, which take into account each group's sample size, are particularly useful in cases where classes are unbalanced. The F-score is a measure that combines precision or positive predictive value (proportion of individuals classified as cases were actually cases) and sensitivity (proportion of true cases correctly classified as such). Balanced accuracy, on the other hand, corresponds to the average accuracy obtained on either class (Brodersen et al., 2010).

2.1.4. Risk of overfitting and possible strategies

Due to the use of multiple nonlinear transformations, deep networks are highly complex models that involve the estimation of a very large number of parameters. This can lead to the model learning particular fluctuations in the training data that are irrelevant

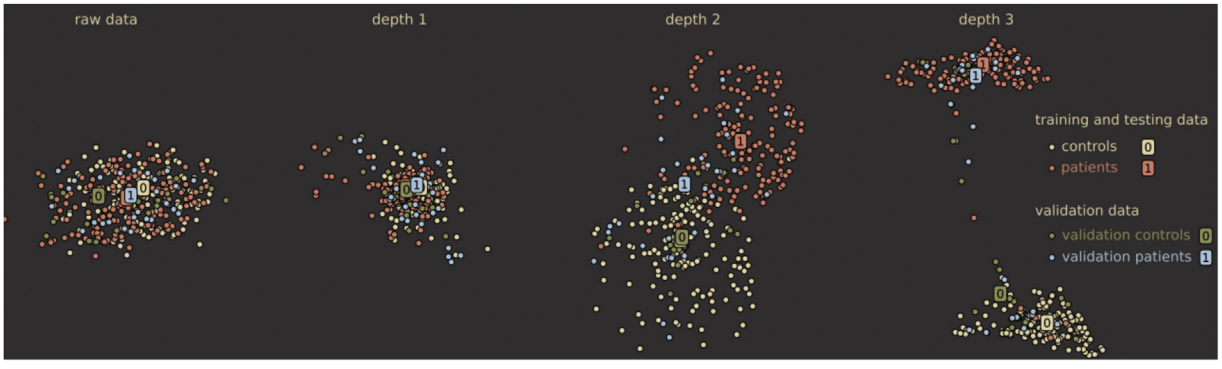

Fig. 2. Effect of the depth of the model. Each dot corresponds to a neuroimage-based data visualized in a two-dimensional map. With more hidden layers, the data becomes more easily separable due to nonlinear transformations along the network (Plis et al., 2014).

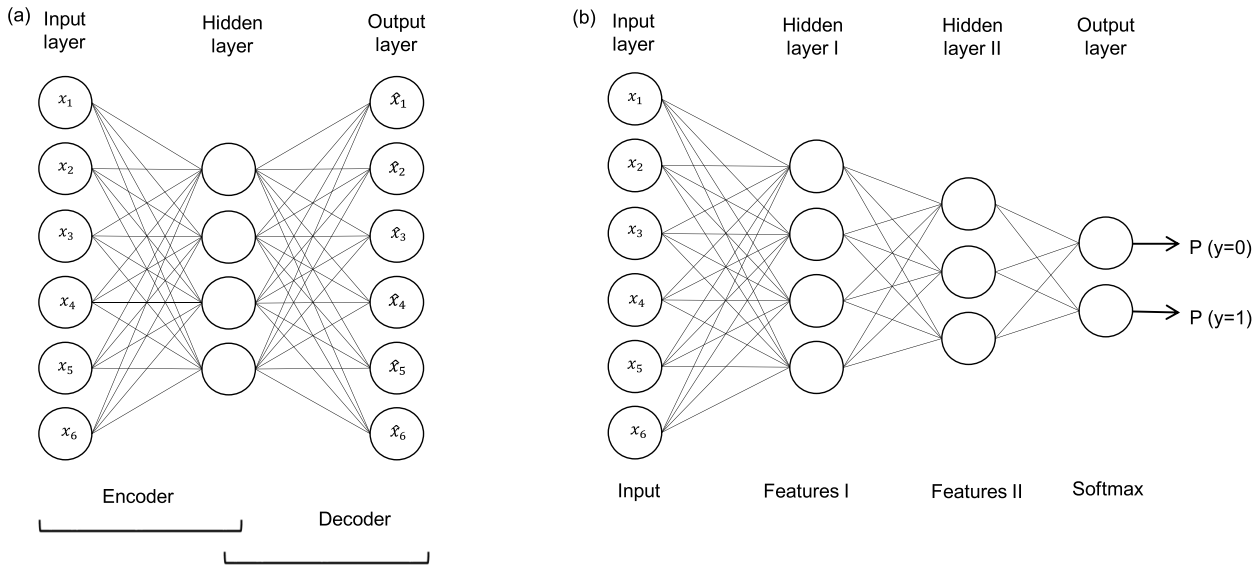

Fig. 3. (a) Shallow or simple autoencoder. In its shallow structure, an autoencoder is comprised of an input layer, that represents the original data (e.g., pixels in an image), one hidden layer that represents the transformed data, and an output layer that reconstructs the original input data. (b) Stacked autoencoder. Two simple autoencoders are stacked with a 2-class softmax classifier as the final layer. From each simple autoencoder, the output layer is discarded, and the hidden layer is used as the input layer for next autoencoder.

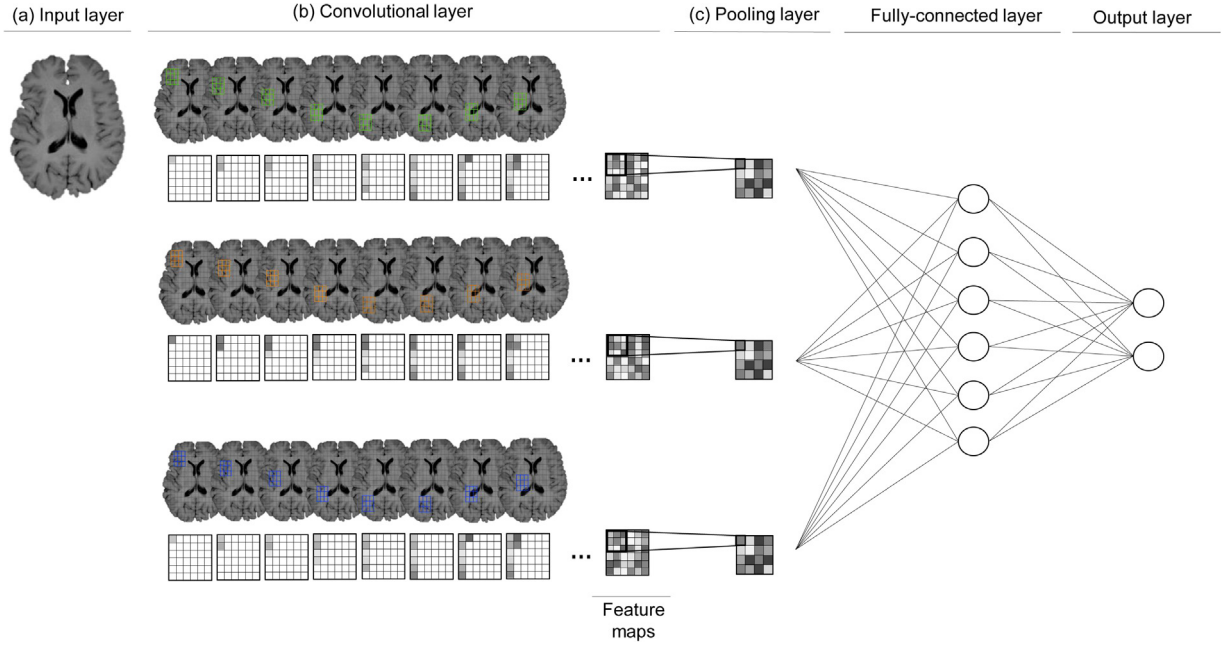

Fig. 4. Generic structure of a CNN. For illustrative purpose, this example only has one layer of each type; a real-world CNN, however, would have several convolutional and pooling layers (usually interpolated) and one fully-connected layer. (a) Input layer. In its simplest way, the data is inputted into the network in such a way that each pixel corresponds to one node in the input layer. (b) Convolutional layer. A 3×3 filter or kernel (in green) is used to multiply the spatially corresponding 3×3 nodes in the image. The resulting weighted sum is then passed through a nonlinear function to derive the output value of one node in the feature map. The repetition of this same operation across all possible receptive fields results in one complete feature map. The same procedure with different kernels (in orange and blue) will result in separate complete feature maps. (c) Pooling layer. The size of each feature map can be reduced by taking the maximum value (or average) from a receptive field in the previous layer. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for the purpose of classification – an issue known as "overfitting". When this happens, the model will perform very well on the training data but will not be able to replicate its performance on unseen data (Srivastava et al., 2014). The risk of overfitting is particularly high in the context of neuroimaging, where the number of data points (e.g. number of voxels) for a subject is much larger than the total number of subjects, resulting in high-dimensional data (Arbabshirani et al., 2016). However, there are a number of strategies that can be used to minimise the risk of overfitting, collectively known as "regularization". A first strategy involves the use of weight decays (e.g., L1 and L2 norms) to penalise models with very high weights. It has been observed that extreme (very low or very high) weight values in a ML model are symptomatic of the model trying to learn the regularities of the data perfectly (Moody et al., 1995). By forcing weights to remain low, the network becomes less dependent on the training data and is able to better generalise to unseen data (Nowlan and Hinton, 1992). A second strategy, known as dropout, consists of temporarily removing

a random number of nodes and their respective incoming and outgoing connections from the network during training. This means that the contribution of dropped-out neurons to the activation of downstream neurons is temporally removed on the forward pass and that any weight updates are not applied to these neurons on the backward pass. The aim of dropout is to extract different sets of features that can independently produce a useful output, thereby allowing higher levels of generalizability (Srivastava et al., 2014).

2.2. Autoencoders

Autoencoders are a special case of feedforward networks which comprise of two main components. The first component, i.e. the "encoder", learns to generate a latent representation of the input data, whereas the second component, i.e. the "decoder", learns to use these learned latent representations to reconstruct the input data as close as possible to the original (Fig. 3a) (Vincent et al., 2010).

Since an autoencoder does not make use of labels, its training is an unsupervised learning process. In its shallow structure, an autoencoder is comprised of three layers: an input layer, one hidden layer and an output layer. The training to perform the input-copying task can be useful to extract meaningful features of the input data. This automatic feature extraction can be performed using an error function (or loss function) that encourages the model encoder to have specific characteristics, such as sparsity of the representation (sparse autoencoders) and robustness to noise (denoising autoencoders). Since autoencoders are automatic features extractors, they can also be stacked to create a deep structure to increase the level of abstraction of learned features. In this case, the network is pre-trained, i.e. each layer is treated as a shallow autoencoder, generating latent representations of the input data. These latent representations are then used as input for the subsequent layers before the full network is fine-tuned using standard supervised learning (Fig. 3b) (Larochelle et al., 2007).

2.3. Deep belief networks

Deep belief networks (DBNs), proposed by Hinton et al. (2006), are technically the first DL models. Similar to stacked autoencoders, DBNs are comprised of stacked shallow feature extractors, known as restricted Boltzmann machines (RBMs). An RBM is composed by only two layers: a visible layer and a hidden layer. Just like autoencoders, RBMs also aim to learn and extract useful features from the data. However, RBMs differ from autoencoders with regards to their training processes. RBMs can be interpreted as a stochastic neural network. Therefore, instead of using deterministic functions and the reconstruction error (like the autoencoders), the RBM uses the maximum-likelihood estimation to find a stochastic representation of the input in its hidden layer (latent features). To do this, RBMs are usually trained using a gradient descent algorithm, with the likelihood gradient being performed by an approximation algorithm known as contrastive divergence (Hinton et al., 2006). Here the input data, stored in the visible layer, are propagated to the hidden layer as in a feedforward network, and the resulting sum of the weighted inputs provides a measure of the neuron activation probability. The activation of hidden neurons can be thought of as the network's internal representation of the data, which is then propagated back to the visible layer in an attempt to reconstruct the input data from the network's internal representation. The network, therefore, learns by adjusting the weights based on the discrepancy between the true and reconstructed data. Similarly to autoencoders, RBMs can be stacked to create a deep network, where the hidden layer representation of one RBM serves as input layer for the following RBM, and the network can learn higher-level features from lower-level ones to arrive at an abstract representation of the data. Furthermore, the neural network corresponding to a trained DBN can be augmented by adding an output layer, where units represent the labels corresponding to the input sample. This results in a standard neural network for classification that can be further trained using supervised learning algorithms.

2.4. Convolutional neural networks

Convolutional neural networks (CNNs) are a special type of feedforward neural networks that were initially designed to process images, and as such are biologically-inspired by the visual cortex (LeCun et al., 1998). In addition to the input and output layers, CNN can comprise of three types of layers: a convolutional layer, a pooling layer, and a fully-connected layer (Fig. 4).

The convolutional layer is organized in several feature maps. Every neuron in a feature map is connected to a fixed set of neurons in a local region of the previous layer – the *receptive field* – in such a way that the whole image is covered ("local connectivity"). Within the same feature map, the connections between each neuron and the corresponding *receptive field* share the same weights, whereas different feature maps use different sets of weights ("weight sharing"). As a result of this architecture, a feature map can be thought of as a "feature detector" that scans the whole image for the same pattern. This pattern is usually known as the kernel. Kernels in a CNN are learned during the training process, as opposed to in SVM, where they are defined a priori. In a network with several convolutional layers, each layer codes for increasingly more abstract features (e.g. lines → edges → eyes → face). The pooling layer simply reduces the number of neurons of the previous convolutional layer. The fully-connected layers are similar to the hidden layers from the conventional MLP where the neurons are connected to all neurons from the previous layer. All combined, the properties of CNN (local connectivity, weight sharing and pooling) result in a significant reduction in the number of parameters, which in turn decreases the likelihood of overfitting, and alleviates computational processing.

3. Review of DL studies of psychiatric or neurological disorders

In order to identify previous applications of DL in neuroimaging studies of psychiatric or neurological disorders, a search was conducted on 1st August 2016 across several databases (PubMed, IEEE Xplore, Scopus and ArXiv) using the following search terms: ("deep learning" OR "deep architecture" OR "artificial neural network" OR "autoencoder" OR "convolutional neural network" OR "deep belief network") AND (neurology OR neurological OR psychiatry OR psychiatric OR diagnosis OR prediction OR prognosis OR outcome) AND (neuroimaging OR MRI OR "Magnetic Resonance Imaging" OR "fMRI" OR "functional Magnetic Resonance Imaging" OR PET OR "Positron emission tomography"). This review did not include EEG studies, although there is some evidence that DL can also be used with this type of data, particularly in epilepsy (Page et al., 2014). The initial search yielded a total of 172 articles. As the next step, we screened and cross-referenced these articles for studies that had applied a deep learning model to neuroimaging data to investigate a psychiatric or neurologic condition; this identified a total of 25 articles which were relevant to our review. We organized these articles as follows: i) *diagnostic studies*, which aimed to classify patients from healthy controls, ii) *studies on conversion to illness*, which used baseline scans from individuals identified as being at high risk of developing a psychiatric or neurologic disorder to predict subsequent transition to the illness, and finally iii) *studies predicting treatment response,* which used baseline scans from individuals with a neurological or psychiatric diagnosis to predict

Table 1
Diagnostic studies.

Authors, year	Sample size	Technique	Features	Previous feature selection	DL architecture	Comparison	Accuracy (%)
Gupta et al. (2013)[a]	AD = 200 MCI = 411 HC = 232	sMRI	WB voxel-level	No	Sparse AE & CNN	HC vs. AD HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	94.7 86.4 88.1 85.0
Payan and Montana (2015)[a]	HC = 755 AD = 755 MCI = 755	sMRI	WB voxel-level	No	Sparse AE & CNN	HC vs. AD HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	95.4 92.1 86.8 89.5
Hosseini-Asl et al. (2016) [a,b]	HC = 70* AD = 70* MCI = 70*	sMRI	WB voxel-level	No	AE & CNN	HC vs. AD HC vs. MCI AD vs. MCI HC vs. AD vs. MCI	97.6 90.8 95.0 89.1
Chen et al. (2015)[a]	HC = 123 AD = 94 MCI = 121	sMRI	WB voxel-level	Yes	SAE	HC vs. AD HC vs. MCI	89.0 81.7
Liu et al. (2015a)[a]	HC = 204 AD = 180 MCI = 374	sMRI	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	82.6 72.0
Gao and Hui (2016)	HC = 117 AD = 51 Lesions = 118	CT	WB voxel-level	No	CNN	HC vs. AD vs. Lesion	87.7
Sarraf and Tofighi (2016)[a]	HC = 15 AD = 28	rsfMRI	WB voxel-level	No	CNN	HC vs. AD	96.9
Suk et al. (2016)[a]	HC = 31 MCI = 31 HC = 25 MCI = 12	rsfMRI rsfMRI	WB region-level WB region-level	Yes Yes	DAE DAE	HC vs. MCI HC vs. MCI	72.6 81.1
Hu et al. (2016)[a]	HC = 52 MCI = 48	rsfMRI	WB region-level	No	SAE	HC vs. MCI	87.5
Han et al. (2015)[a]	HC = nr AD = nr	rsfMRI	WB voxel-level	No	AE & CNN	HC vs. AD	80.0
Liu et al. (2015a)[a]	HC = 77 AD = 85 MCI = 169	sMRI & PET	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	91.4 82.1
Suk et al. (2014)[a]	HC = 101 AD = 93 MCI = 204	sMRI & PET	WB region-level	Yes	DBM	HC vs. AD HC vs. MCI	94.9 80.6
Liu et al. (2014)[a]	HC = 77 AD = 65 MCI = 169	sMRI & PET	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI	87.8 76.9
Suk et al. (2015b)[a]	HC = 52 AD = 51 MCI = 99 HC = 229 AD = 198 MCI = 403	sMRI & PET & CSF sMRI & PET & CSF	WB region-level WB region-level	Yes Yes	DW-S2 MTL DW-S2 MTL	HC vs. AD HC vs. MCI HC vs. AD vs. MCI HC vs. AD HC vs. MCI HC vs. AD vs. MCI	95.1 80.1 62.9 90.3 70.9 57.7
Liu et al. (2015b)[a]	HC = 77 AD = 85 MCI = 169	sMRI & PET & MMSE	WB region-level	Yes	SAE	HC vs. AD HC vs. AD vs. MCI	90.1 59.2
Suk et al. (2015a)[a]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	SAE	HC vs. AD HC vs. MCI AD vs. MCI	98.8 90.7 83.7
Li et al. (2014)[a]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	MLP	HC vs. AD HC vs. MCI	91.4 77.4
Suk and Shen (2013)[a]	HC = 52 AD = 51 MCI = 99	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	No	SAE	HC vs. AD HC vs. MCI	95.9 85.0
Han et al. (2015)[c]	HC = nr ADHD = nr	rsfMRI	WB voxel-level	No	AE & CNN	HC vs. ADHD	65.0
Deshpande et al. (2015)[c]	HC = 744 ADHD-C = 260 ADHD-I = 173	rsfMRI	WB region-level	Yes	FCC	HC vs. ADHD-C HC vs. ADHD-I ADHD-C vs. ADHD-I	~90.0 ~90.0 95.0

Table 1 (*Continued*)

Authors, year	Sample size	Technique	Features	Previous feature selection	DL architecture	Comparison	Accuracy (%)
Kuang et al. (2014)[c]	HC = 69 to 110 ADHD-C = 16 to 95 ADHD-I = 2 to 5 ADHD-H = 1 to 50	rsfMRI	ROI (PFC) ROI (VC) ROI (CC)	Yes	DBN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H HC vs. ADHD-C vs. ADHD-I vs. ADHD-H HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	37.4 to 71.8[***] 34.4 to 68.8[***] 37.1 to 72.7[***]
Kuang and He (2014)[c]	HC = 42 to 95 ADHD-C = 0 to 77 ADHD-I = 0 to 44 ADHD-H = 0 to 6	rsfMRI	ROI (PFC)	Yes	DBN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	44.4 to 80.9[***]
Hao et al. (2015)[c]	HC = 69 to 110 ADHD-C = 16 to 95 ADHD-I = 2 to 5 ADHD-H = 1 to 50	rsfMRI	ROI (PFC, VC, SSC and CC combined)	Yes	DBaN	HC vs. ADHD-C vs. ADHD-I vs. ADHD-H	48.9 to 72.7[***]
Plis et al. (2014)	HC = 191 SZ and FEP = 198	sMRI	WB voxel-level	No	DBN	HC vs. SZ	91[**]
Kim et al. (2016)[d]	HC = 50 SZ = 50	rsfMRI	WB region-level	Yes	SAE	HC vs. SZ	85.8
Munsell et al. (2015)	HC = 48 TLE = 70	DTI	WB region-level	No	SAE	HC vs. TLE	69.0
Yang et al. (2014)	HC = 31 SCA2 = 4 SCA6 = 27 AT = 18	sMRI	ROI (Cerebellum)	No	SAE	HC vs. SCA2 vs. SCA6 vs. AT	86.3

[a] ADNI dataset.
[b] CADDementia dataset.
[c] ADHD-200 dataset.
[d] COBRE dataset.
[*] Sample sizes for the fine-tuning stage only (pre-training included an additional 386 samples).
[**] F-score.
[***] Range of accuracies obtain from the different datasets used; HC, healthy controls; SZ, schizophrenia, FEP, first episode psychosis; ADHD, attention deficit/hyperactive disorder; ADHD-C, attention-deficit/hyperactive disorder combine subtype; ADHD-I, attention-deficit/hyperactive disorder inattentive subtype; ADHD-H, attention-deficit/hyperactive disorder hyperactive subtype; SCA2, spinocerebellar ataxia type 2; SCA6, spinocerebellar ataxia type 6; AT, ataxia-telangiectasia; TLE, temporal lobe epilepsy; AD, Alzheimer's disease; MCI, mild cognitive impairment; CC, cingulate cortex; VC, visual cortex, PFC, pre-frontal cortex; SSC, somatosensory cortex; sMRI, structural MRI; rsfMRI, resting-state functional MRI; CT, computed tomography; PET, Positron emission tomography; DTI, diffusion tensor imaging; CSF, cerebrospinal fluid; MMSE, mini mental state examination; ADASCog, Alzheimer's Disease Assessment Scale's cognitive subscale; AE, autoencoder, SAE, stacked autoencoder; FCC, fully-connected cascade; DBN, deep belief network, DBaN, deep Bayesian network; CNN, convolutional neural network; DAE, deep autoencoder; DBM, deep Boltzman machine; DW-S2 MTL, deep weighted subclass-based sparse multi-task learning; MLP, multilayer perceptron; nr, not reported.

subsequent treatment response. These studies are summarised in Tables 1, 2 and 3 which provide the following information: sample size; type of data used as input; whether a whole brain (WB) or region of interest (ROI) approach was used; whether the information inputted into the model comprised of voxel or region-level features; whether feature selection was or was not used before inputting the data into the model; general type of DL architecture; diagnostic groups being investigated; and accuracy. Whenever performed, we also report the accuracies obtained for multiclass classifications, which involve discriminating between more than two classes (e.g. healthy controls vs. mild cognitive impairment vs. Alzheimer's disease).

3.1. Diagnostic studies

Studies using DL to classify psychiatric or neurological patients from healthy individuals have used a range of neuroimaging modalities including structural MRI (sMRI), resting-state fMRI (rsfMRI), positron emission tomography (PET) and a combination of different modalities (multimodal studies) (see Table 1). From Table 1 it can be seen that the vast majority of these studies were carried out in Alzheimer's disease (AD) and its prodromal stage, mild cognitive impairment (MCI). In addition, a smaller number of studies examined psychosis, attention deficit/hyperactivity disorder (ADHD), cerebellar ataxia and temporal lobe epilepsy (TLE). Within each diagnostic category, we first give an overview of the studies that have used a single neuroimaging modality, followed by studies that employed a multimodal approach and, finally, studies that have combined neuroimaging and clinical data within a single classifier.

3.1.1. Mild Cognitive Impairment and Alzheimer Dementia

In one of the first studies using DL in AD and MCI, Gupta et al. (2013) argued that, since (i) natural images and brain imaging have similar, and therefore interchangeable, low-level features (e.g. lines and corners) and (ii) natural images, contrary to neuroimaging, are abundant, then natural images could be used to learn low level features which could then be used to identify lesions along the surface and ventricles of the brain. This process, whereby the features

Table 2
Conversion to illness.

Authors, year	Sample size	Technique	WB voxel-level/WB region-level/ROI	Previous feature selection	DL architecture	Comparison	Accuracy (%)
Liu et al. (2015a)[a]	HC = 204 AD = 180 MCI-C = 160 MCI-NC = 214	sMRI	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	46.3
Suk et al. (2014)[a]	MCI-C = 76 MCI-NC = 128	sMRI & PET	WB region-level	Yes	DBM	MCI-NC vs MCI-C	71.6
Liu et al. (2015a)[a]	HC = 77 AD = 85 MCI-C = 67 MCI-NC = 102	sMRI & PET	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	53.8
Liu et al. (2014)[a]	HC = 77 AD = 65 MCI-C = 67 MCI-NC = 102	sMRI & PET	WB region-level	Yes	SAE	AD vs MCI-C vs MCI-NC vs HC	47.4
Suk et al. (2015b)[a]	MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	MCI-NC vs MCI-C	74.2
	AD = 51 HC = 52 MCI-C = 167 MCI-NC = 236					AD vs MCI-C vs MCI-NC vs HC	53.7
		sMRI & PET & CSF	WB region-level	Yes	DW-S2 MTL	MCI-NC vs MCI-C	73.9
						AD vs MCI-C vs MCI-NC vs HC	47.8
Li et al. (2014)[a]	HC = 52 AD = 198 MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	MLP	MCI-NC vs MCI-C	57.4
Suk and Shen (2013)[a]	MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	No	SAE	MCI-NC vs MCI-C	75.8
Suk et al. (2015a)[a]	MCI-C = 43 MCI-NC = 56	sMRI & PET & CSF & MMSE & ADASCog	WB region-level	Yes	SAE	MCI-NC vs MCI-C	83.3

[a] ADNI dataset; HC, healthy controls; AD, Alzheimer's disease; MCI-NC, mild cognitive impairment non-converters; MCI-C, mild cognitive impairment converters; sMRI, structural MRI; PET, Positron Emission Tomography; CSF, cerebrospinal fluid; MMSE, mini mental state examination; ADASCog, Alzheimer's Disease Assessment Scale's cognitive subscale; SAE, stacked autoencoder; DBM, deep Boltzmann machine; DW-S2 MTL, deep weighted subclass-based sparse multi-task learning; MLP, multilayer perceptron.

Table 3
Treatment outcome.

Authors, year	Sample size	Technique	WB voxel-level/WB region-level/ROI	Previous feature selection	DL architecture	Comparison	Accuracy (%)
Munsell et al. (2015)	TLEns = 41 TLEs = 29	DTI	WB region-level	No	SAE	TLEns vs TLEs	57.0

HC, healthy controls; TLE-ns, temporal lobe epilepsy without seizures; TLE-ns, temporal lobe epilepsy with seizures; DTI, diffusion tensor imaging.

learned in one set of data are used to solve a problem in another set of data, is known as "transfer learning". Based on this premise, the authors pre-trained a sparse autoencoder to learn features from natural images, which were then applied to structural MRI data via a CNN, achieving a classification accuracy of 94.7% for AD versus controls, 86.4% for MCI versus controls and 88.1% for AD versus MCI. Consistent with the authors' hypothesis, this method outperformed the one where the learned features were extracted from the neuroimaging data (93.8%, 83.3% and 86.3% for the same comparisons, respectively). However, a few years later and using a similar approach, Payan and Montana (2015) found comparable classification accuracies using features that were learned from the structural MRI data itself. This could potentially be explained by the fact that Payan and Montana (2015) used a much larger sample, as well as by the fact that authors used 3D brain images, as opposed to 2D, which possibly contain more useful patterns for classification. Indeed, Payan and Montana (2015) reported that, in general, the models based on 3D outperformed those based on 2D brain images (AD vs. HC (2D/3D) = 95.4%/95.4%; AD vs. MCI (2D/3D) = 82.2%/86.8%; MCI vs. HC (2D/3D) = 90.1%/92.1%). The best accuracy (97.6%) from single modality studies came from Hosseini-Asl et al. (2016), who also used transfer learning. Instead of extracting features from natural images and then fine-tuning the model on Alzheimer's patients and controls, as seen in Gupta et al. (2013); Hosseini-Asl et al. (2016) used one Alzheimer's dataset for pre-training and another independent Alzheimer's dataset to fine-tune the model. By performing the pre-training on an Alzheimer's dataset, this approach allowed for the network to extract generic features related to AD biomarkers, such as the ventricular size, hippocampus shape, and cortical thickness as opposed to more generic low-level features as in Gupta et al. (2013). By using two independent samples during the complete learning process, the final learned features for classification

are much less dataset-specific, and should therefore be more generalizable. The final model's architecture was also deeper than in previous studies, which probably also contributed to the high accuracy. Taken collectively, these studies suggest that the application of DL to structural MRI data allows the classification of individuals with AD and MCI with high levels of accuracy. Consistent with the increasing popularity of CNN models, studies that have applied either CNN or a combination of AE and CNN have shown better performances compared to those using only AE, although it should be noted that the former group of studies tended to have larger samples than the latter group. In addition, and similar to the trend reported in computer vision competitions and research, the best performances were obtained by the deepest CNN models.

Studies of AD and MCI using resting-state imaging have also achieved promising results. For example, Han et al. (2015) designed a hierarchical convolutional sparse autoencoder (HCSAE), which essentially extracts the most discriminating features from the resting-state data and encodes them in a convolutional manner. This particular arrangement allows for the extraction of the most useful information while conserving abundant detail. The final model classified AD and controls with an 80.0% accuracy and significantly outperformed SVM, which only yielded an accuracy of 50% (Fig. 4). While this is a promising result, the model assumed that functional networks were statistic over time – an assumption which underlies the vast majority of ML applications to resting-state neuroimaging data. However, recent studies have shown that the network-level functional organization of the brain is dynamic rather than static (Hutchison et al., 2013). Suk et al. (2016) have addressed this issue by developing an approach which classifies people with MCI and healthy controls using a deep autoencoder to extract hierarchical nonlinear relations among brain regions, whilst modelling the inherent functional dynamics of resting-state data. This was also one of the few studies in which the same DL model was tested against and surpassed other competing models in two independent datasets (72.6% for dataset 1 and 80.0% for dataset 2), thus providing evidence of replicability, a crucial feature for diagnostic tools. In line with the studies using structural imaging, the best performance for the classification of AD patients with resting-state data was also obtained by a CNN model with an accuracy of 96.9% (Sarraf and Tofighi, 2016). These studies provide initial evidence that brain activity at resting state can be useful in identifying MCI and AD patients. We note that, compared to the performances obtained from structural data, DL models applied to functional data seem to perform worse. This discrepancy could be explained by the substantial difference in sample size between the two types of studies – while the *smallest* study using structural data included 140 subjects (Hosseini-Asl et al., 2016) the *largest* study using functional data included 62 subjects (Suk et al., 2016).

With regards to multimodal studies, Liu et al. (2014) applied a stacked autoencoder (SAE) to structural and PET data and successfully distinguished AD and MCI from controls with an accuracy of 87.8% and 76.9%, respectively. Using a very similar dataset, the same team (Liu et al., 2015a) achieved a better performance by designing a model where the hidden layers were able to infer the correlations between sMRI and PET, thus better capturing the synergy between the two modalities. This model classified AD and MCI against controls with an accuracy of 91.4% and 82.1%, respectively. Interestingly, the application of the same model to a structural data alone resulted in less impressive accuracies of 82.6% and 72% for AD and MCI, respectively. This discrepancy suggests that the integration of structural and functional data may improve classification accuracy. However, this conclusion should be drawn with great caution since that the authors did not report classification accuracy for PET data alone.

Finally, four studies have tried combining neuroimaging data with clinical information to build a more robust classification model. For example, Suk and Shen (2013) used a SAE to extract latent features from neuroimaging data (sMRI, PET and CSF), which were then used to predict clinical data (measured using the Mini-Mental State Examination – MMSE – and Alzheimer's Disease Assessment Scale's cognitive subscale – ADAS-cog) and class labels. As the final step, the resulting learned features were used to classify AD and MCI from healthy individuals with an accuracy of 95.9% and 85.0%, respectively. Notably, two more studies (Li et al., 2014; Suk et al., 2015a) that have used the same exact sample (taken from the publicly available dataset ADNI; Alzheimer's Disease Neuroimaging Initiative) and the same types of data (sMRI, PET, CSF, MMSE and ADAS-cog) have also reported high accuracies for both AD and MCI despite using different implementations of DL. In general, studies combining clinical with neuroimaging data have, in general, reported higher accuracies than studies using single modality or multiple neuroimaging modalities. This is in line with previous studies using conventional ML methods (e.g. Willette et al., 2014; Moradi et al., 2015; Zhang and Shen, 2012) and highlights the usefulness of adding clinical information in the classification of AD and its prodromal phase.

3.1.2. Attention-deficit/hyperactive disorder

With regards to attention-deficit/hyperactivity disorder (ADHD), all five studies included here have used resting-state neuroimaging data. For example, Deshpande et al. (2015) applied a fully connected cascade artificial neural network – a variation of the multilayer perceptron – to functional connectivity from ADHD and healthy controls. The model successfully distinguished between the inattentive and combined subtypes from healthy controls with an accuracy of 90% for both comparisons, while the two subtypes were discriminated with an accuracy of 95%. Connections between frontal areas and the cerebellum were identified as the most discriminating features. There is also evidence that healthy children and children diagnosed with three different ADHD subtypes (inattentive, hyperactive and combined) can be distinguished in one single model using a multiclass approach, without the need to perform binary classifications between healthy controls and each ADHD subtypes. This evidence comes from three studies that have used data from different sites taken from the ADHD-200 consortium, a data-sharing platform aimed at understanding the neural basis of ADHD (Milham et al., 2012). Kuang et al. (2014) attempted to discriminate between healthy controls and ADHD subtypes (inattentive, hyperactive and combined) using data acquired from three different sites. Rather than looking at the whole brain, the authors first parcellated the brain and trained different DBNs for each brain area to examine which part of the brain best discriminated ADHD (regardless of subtypes) from healthy controls. A 4-way DBN was then performed for the each best discriminating area – prefrontal (PFC), cingulate (CC) and visual (VC) cortex – in each one of the three datasets separately (dataset 1: PFC = 37.4%, CC = 37.1%, VC = 34.4%; dataset 2: PFC = 54.0%, CC = 54.0%, VC = 51.2%; dataset 3: PFC = 71.8%, CC = 72.7%, VC = 68.8%). Kuang and He (2014) partially replicated these findings by applying the same DL approach to functional measures of the prefrontal cortex; this allowed a 4-way classification accuracy of 44.4%, 55.6% and 80.9% in three independent samples from the ADHD-200 consortium. Finally, Hao et al. (2015) identified the most discriminating areas – prefrontal, cingulate, somatosensory and visual cortex – and then combined them within a single model. The resulting input data were put through a deep Bayesian network (DBaN), where a DBN was used to reduce the dimensionality of the data and a Bayesian network was used to extract the relationships between the data. The resulting model achieved a 4-way classification accuracy of 48.8%, 54.0% and 72.7% for three independent samples also taken from the ADHD-200 consortium. These three studies suggest that DL can be used to

solve multiclass classifications problems, as all performances were well above chance level (25% for a classification with 4 classes). In addition, these studies suggest that DL can extract meaningful information from patterns of brain functioning to classify ADHD from controls and, more notably, to differentiate between ADHD subtypes. Nevertheless, we note that all four studies conducted in ADHD had unbalanced sample sizes between classes. For example, in Kuang et al. (2014), there were just between 2 and 5 children in the Inattentive subtype within each site, while the number of healthy children ranged from 69 to 110 per site. Similarly, each site in Kuang and He (2014) did not include any participants on at least one ADHD subtype which may have introduced a bias in the 4-way classification performed across all sites. With the exception of Hao et al. (2015) which reported sensitivity and specificity, all studies assessed model performance by estimating the overall accuracy. This metric is simply the proportion of participants correctly identified, and therefore does not take the unbalance between classes into account; this means that it is possible to have a good overall accuracy even if several participants from a class are misclassified (or even if all participants from a class are misclassified if the sample size for that class is very small compared to the total sample size). Therefore, given the highly imbalanced sample sizes, the possibility that the performances reported in these studies are inflated cannot be ruled out. This possibility is supported by the observation of much lower sensitivities (43.9%, 22.9% and 55.6% for each site) than specificities (68.8%, 87.7% and 83.0%), in Hao et al. (2015).

3.1.3. Psychosis

With respect to psychosis, two studies have been performed with promising results. Using structural MRI data from four independent studies, Plis et al. (2014) applied a DBN to the original pre-processed images obtaining an impressive F-score of 91%. While this was a highly promising result, the patients group included both first episode and chronic schizophrenia patients, which could have diluted the models' performance. More recently, Kim et al. (2016) extracted functional connectivity patterns obtained from resting-state functional MRI of individuals diagnosed with schizophrenia and healthy controls and performed a series of experiments with an SAE-based model, in which different hyperparameters were tested. The proposed model consisted of an SAE with weight sparsity control, i.e. only a random selection of neurons in a given layer was activated, that classified schizophrenia patients and controls with an accuracy of 85.5%, outperforming SVM by a margin of 8.1%. Consistent with the literature on brain functional abnormalities in schizophrenia (Kühn and Jürgen, 2013; van der Meer et al., 2010), the most relevant features for the classification were the functional connectivity between the thalamus and the cerebellum, the frontal and temporal areas and between the precuneus/posterior cingulate cortex and the striatum. Despite this encouraging result, the sample sizes for each class were modest (50 for each group) and, therefore, it is not clear how well these findings will generalise to a different sample. Nevertheless, both studies suggest that DL can effectively classify psychosis patients on the basis of neuroanatomical and neurofunctional information. Despite the evidence that structural and functional data provide complementary information on the neural basis of psychosis (Cabral et al., 2016; Radua et al., 2012; Schultz et al., 2012), to date there have been no DL studies using a multimodal approach in psychosis. In addition, despite the evidence that psychosis, similar to AD, is preceded by a prodromal stage (Yung et al., 2005), there have been no studies applying DL to neuroimaging data to classify individuals at high risk of developing psychosis from healthy controls or distinguishing between high risk individuals who will and will not develop the illness.

3.1.4. Temporal lobe epilepsy

One study examined the potential of DL to classify healthy individuals and patients diagnosed with temporal lobe epilepsy (TLE) from diffusion-weighted images (DWI) (Munsell et al., 2015). A stacked autoencoder was used to extract meaningful features from patients' connectome while SVM was chosen as the classifier. Deep learning was suggested as an attractive ML alternative because it is capable of encoding latent, nonlinear relationships in high dimension data. This combination yielded a relatively modest accuracy of 69%. In addition, this model was outperformed by another approach where features were extracted using a well-known linear automated method (ElasticNet) instead, which achieved an accuracy of 80%. This discrepancy in favour of the second model could potentially be explained by the absence of any form of regularizers in the first model. Given the high complexity resulting from the numerous parameters to be estimated, DL models are more prone to overfitting (high performance on the training data while performing poorly on unseen data) than conventional ML approaches. One standard solution, that the authors did not use, is to address this issue is by tuning the level of model complexity and penalizing highly intricate ones in order to have better generalizing models.

3.1.5. Cerebellar ataxia

One study was conducted in cerebellar ataxia (CA), a neurodegenerative disorder that affects mainly the cerebellum, with multiple genetics variations each with its characteristic pattern of anatomical degeneration. Yang et al. (2014) applied a stacked AE to T1-weighted images of the cerebellum taken from healthy controls and individuals suffering from three CA subtypes: spinocerebellar ataxia type 2 (SCA2), spinocerebellar ataxia type 6 (SCA6) or ataxia-telangiectasia (AT). The proposed method classified the four groups with an accuracy of 86.3%, an impressive result for a 4-way classification. However, the confusion matrix reported by the authors indicates that no case with the SCA2 subtype was correctly classified. Because the sample size of this group (only four participants) contributed very little for the total sample size (80), it is still possible to misclassify all its cases and achieve a low error rate. In such cases, a high accuracy can be misleading, as it may reflect an overestimation of the algorithm's performance (Arbabshirani et al., 2016). Balanced accuracy, for example, is a potentially useful alternative as it calculates the average of correct predictions of each class individually (Alberg et al., 2004).

In short, since the first study published in 2013, there is already preliminary evidence that DL allows the accurate classification of a range of neurologic and psychiatric disorders, by extracting discriminating features from either single or multimodal imaging as well as other types of data such as clinical and cognitive information.

3.2. Conversion to illness

3.2.1. From Mild Cognitive Impairment to Alzheimer Dementia

A total of 8 studies have attempted to predict transition to illness using neuroimaging data, and all of them have focussed on the transition from MCI to AD (Table 2). With one exception (Liu et al., 2015a), all studies used a multimodality approach, with three of them also including clinical measures in the prognostic model. The highest accuracy (83.3%), was achieved by a model which included sMRI, PET, CSF and two clinical measures: the MMSE and the ADAS-cog (Suk et al., 2015a). Interestingly, the lowest performance (57.4%) resulted from a model which used the same input data (sMRI, PET, CSF, MMSE and ADASCog) and a similar sample size (Li et al., 2014). However, the two studies differed on the DL approach, with the former employing a semi-supervised approach with a multilayer perceptron pretrained using a stacked sparse autoencoder, and the latter using a pure supervised approach.

These findings highlight the potential impact of the DL architecture on performance, although we cannot exclude the contribution of other sample-specific factors to the results (e.g. recruitment criteria). Overall, this initial sample of studies suggests that individuals diagnosed with MCI who later convert to dementia can be identified using cutting-edge DL methods. Although, in general, accuracies are not as high as when classifying AD or MCI from healthy controls, this is not surprising since brain differences as well as clinical and cognitive symptoms between those identified as being at risk who do and do not develop a disorder are likely to be subtle. In addition to these encouraging results, the suitability of DL to multiclass classification means this analytical approach can easily be employed to examine the biomarkers of different stages of the illness. Four studies have taken advantage of this by conducting 4-way classifications to discriminate between no eminent risk of AD (healthy controls), individuals in the prodromal stage who did not (MCI-C) and did develop dementia (MCI-C) and established Alzheimer's (AD). Accuracies ranged from 46.3% to 53.8%. By using a deep Boltzmann machine to extract features from structural MRI and PET images, Liu et al. (2015a) classified the four groups with an overall accuracy of 53.8%. Suk et al. (2015b) examined the replicability of a DL approach known as deep weighted subclass-based sparse multi-task learning (DW-S2 MTL) in two different datasets, considering both binary and multi-way comparisons. The proposed model, specifically designed to mitigate the effect of less useful features for classification, showed a comparable performance for both binary (74.2% vs. 73.9%) and 4-way (53.7% vs. 47.8%) classifications, thus suggesting good replicability. Taken collectively, these studies provide initial evidence that DL methods could be used to discriminate amongst different stages of illness – a common challenge in standard clinical settings.

3.3. Treatment outcome

Prediction of response to treatment is a research area of high clinical interest. In several psychiatric and neurological disorders, a better understanding of why some patients benefit from a certain treatment whereas others do not, could help clinicians make more-effective treatment decisions and improve long-term clinical outcomes (Mechelli et al., 2015). However, so far, only one study has used DL to predict clinical response to treatment (Table 3). Munsell et al. (2015) attempted to develop an algorithm that distinguished between patients with TLE who did and did not benefit from surgical treatment. This was implemented using a stacked autoencoder to extract meaningful features from the connectome of patients who were then classified using SVM. This model, however, yielded a low accuracy of 57%. For comparison, the author investigated another option where features were extracted with an alternative linear approach instead of an autoencoder. This second model resulted in a higher accuracy of 70%. Again, this discrepancy in favour of the second model could potentially be explained by the absence of any form of regularizers in the first model. This model comprised 4 layers, resulting in a high number of weights to be estimated which, together with a modest sample size (41 patients without seizures and 29 with seizures after treatment), might have resulted in overfiting.

3.4. How does DL compare to a traditional machine learning approach?

A total of twenty-five studies included in this review compared a DL model against a kernel-based model (SVM or MKL) in order to elucidate how DL compares to a more conventional ML approach. The results of these comparisons are shown in Fig. 5. It can be seen that, for the majority of studies, DL showed improved performance compared to SVM. Given the small sample of studies, it is difficult to identify specific characteristics of the studies associated with greater or smaller improvement in performance following the implementation of DL. However, a margin favouring DL studies appears to be more evident in studies that have integrated different modalities with cognitive and/or clinical data (Fig. 6). This anecdotal observation is consistent with the notion that DL is a powerful tool for detecting abstract relations within the data, especially between different types of data that are likely to be associated in complex ways, such as neuroimaging and clinical/cognitive information (Plis et al., 2014).

Since DL requires a large number of observations to learn increasingly complex patterns compared to conventional ML methods, one would expect to find a greater difference between the two methods as sample size increases. However, the effect of sample size on the difference in performance is unclear, possibly due to the small number of studies currently available. There is a minority of studies where SVM/MKL matched or even outperformed the proposed DL model. Amongst these, Munsell et al. (2015) reported the largest margin favouring SVM. However, this article had one of the smallest sample sizes (118 for the diagnostic comparison and 70 for the treatment outcome comparison) while employing one of the deepest networks with 5 layers. Notably, out of all the studies comparing the two approaches, Munsell et al. (2015) was the only one that did not make any formal attempt to prevent overfitting of the DL model, for example through the use of regularization. We note that susceptibility to overfitting becomes more pronounced when deeper and thus more complex networks are used, as in the study by Munsell et al. (2015), due to the higher number of weights to be estimated (Srivastava et al., 2014). Therefore, we speculate that the use of small sample sizes, coupled with the high-dimensionality of the data (i.e. when the number of variables highly exceeds the number of participants), may have increased the risk of overfitting in this study.

4. Discussion

ML has been gaining considerable attention in the neuroimaging community due to its advantages over traditional analytical methods based on mass-univariate statistics. In particular, ML methods take the inter-correlation between regions into account, while mass-univariate methods operate under the assumption that different regions act independently. In addition, ML methods can be used to make inferences at the single-subject level – a critical difference with mass-univariate analytical methods that are only sensitive to differences at group-level. DL is a type of ML which is increasingly used in neuroimaging after leading to major scientific advances in the areas of speech recognition, computer vision and natural language processing by significantly outperforming other state-of-the-art classification methods (Krizhevsky et al., 2012; Le et al., 2012). There are two main characteristics that distinguish DL from conventional ML methods: first, DL is capable of learning features from the raw data without the requirement for *a priori* feature selection, resulting in a more objective or less bias-prone process; second, DL uses a hierarchy of nonlinear transformations, which make this approach ideally suited for detecting complex, scattered and subtle patterns in the data. Given its ability to detect abstract patterns from the data, DL can be considered a promising tool in neuroimaging, as most brain-based disorders are characterised by a scattered and diffused pattern of neuroanatomical and neurofunctional alterations (Plis et al., 2014). In previous sections of this review, we have described the most common DL architectures and have provided an overview of the studies that have applied DL to neuroimaging data to investigate psychiatric and neurological disorders. In this final section, we discuss the main themes that have emerged from the review of these studies. These will include (i)

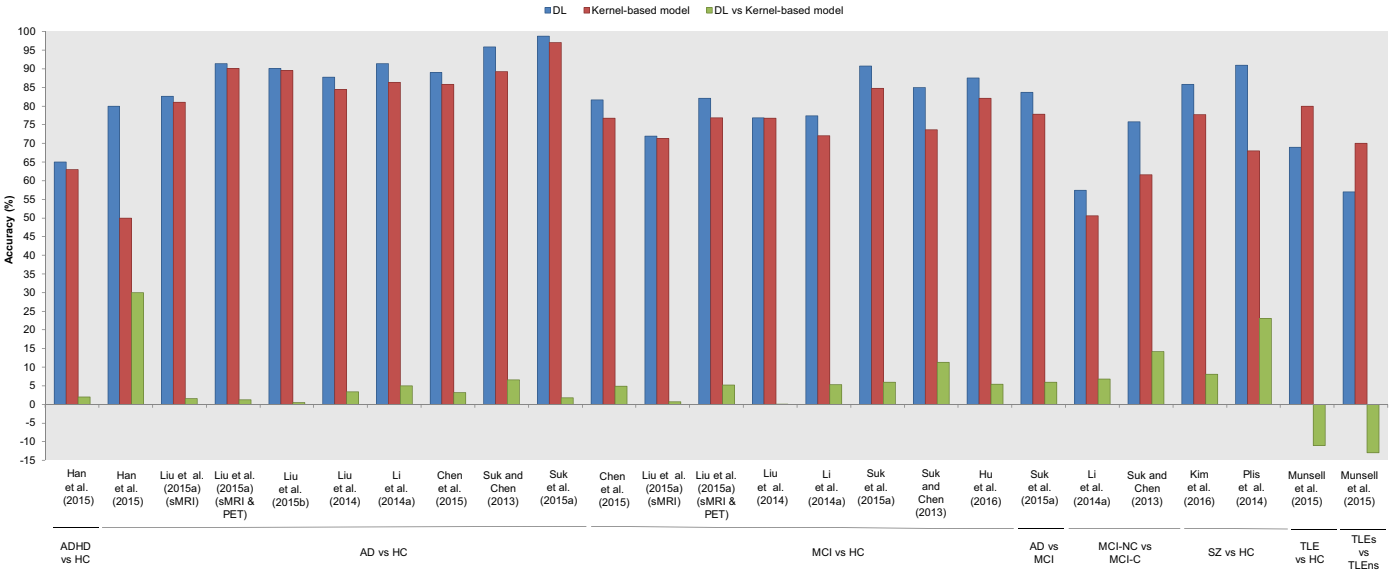

Fig. 5. Results of studies comparing DL and kernel-based models. The graph shows the accuracies (F-score for Plis et al., 2014) for DL models (blue), kernel-based models (red) and the difference between the two (green). HC, healthy controls; ADHD, attention deficit and hyperactive disorder; AD, Alzheimer's disease; MCI, mild cognitive impairment; MCI-NC, mild cognitive impairment non-converters; MCI-C, mild cognitive impairment converters; SZ, schizophrenia; TLE, temporal lobe epilepsy; TLEs, temporal lobe epilepsy with seizures after treatment; TLEns, temporal lobe epilepsy without seizures after treatment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

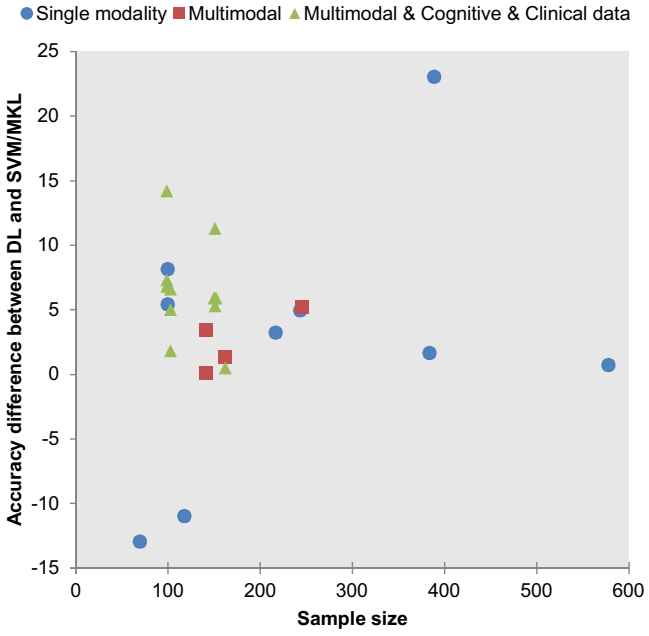

Fig. 6. Difference in performance of DL against kernel-based methods for single modality, multimodal as well as for multimodal with cognitive/clinical data studies, according to sample size.

consistencies and inconsistencies in the existing literature (ii) the promise of CNNs, (iii) the issue of multiclass classification, (iv) how DL performs compared with conventional ML methods, (v) interpretability of DL in neuroimaging, (vi) the challenge of overfitting and (vii) technical expertise and computational requirements. We conclude by discussing possible directions for future research.

4.1. Main conclusions from the existing literature

The majority of published studies have been conducted in patients with MCI and/or AD; this may be explained by the availability of ADNI, a very large open-source dataset including thousands of patients, to the neuroimaging community (Mueller et al., 2005a, 2005b). However, studies have also been conducted in other disorders including ADHD, psychosis, TLE and cerebellar ataxia. Taken collectively, the findings published so far suggest that DL can be applied to neuroimaging data, including both structural and functional modalities, to classify diagnostic groups from healthy individuals. Indeed, the performance of the classifiers has been consistently high, with several studies reporting accuracies above 95% for binary classifications between patients and controls (Deshpande et al., 2015; Hosseini-Asl et al., 2016; Payan and Montana, 2015; Sarraf and Tofighi, 2016; Suk and Shen, 2013; Suk et al., 2015a; Suk et al., 2015b). Nevertheless, the application of a supervised model for diagnostic classification is arguably circular: since diagnostic labels in the training and testing datasets are predetermined through clinical examination, logic dictates that a perfect performance from an ML algorithm will simply mimic clinical assessment. Being able to predict a future diagnosis, or anticipate who will and will not benefit from a certain treatment, are questions of greater translational value in clinical practice. A total of 8 studies have applied DL to neuroimaging data acquired from individuals with MCI to predict subsequent transition to AD with promising results. For example, Suk et al. (2015a) successfully predicted conversion from MCI to AD with 83.3% accuracy, after combining structural MRI and PET data. However, no studies have yet examined transition to illness in other psychiatric disorders with a prodromal phase, such as psychosis, even though we know that it is possible to distinguish between converters and non-converters using conventional ML (Zarogianni et al., 2013; Pettersson-Yeo et al., 2013; Valli et al., 2016). To our knowledge only one study has used DL to predict treatment outcome. Munsell et al. (2015) achieved an accuracy of 57% when classifying TLE patients who did and did not suffer from seizures after surgical intervention. As discussed earlier, however, this modest result could potentially be explained by the absence of formal strategies to avoid overfitting of the DL model.

DL is a very flexible approach, meaning that is it possible to combine different architectures and manipulate a range of hyperparameters within the same model. In addition, the vast majority

of existing studies have been published in the last 2 years, and therefore the field of DL applied to neuroimaging of brain-disorders should be considered still at a very early stage. Possibly as a result of this combination of flexibility and novelty, the methodology of the studies reviewed in this article varied considerably. For example, some studies employed a whole-brain approach whereas others focussed on a subset of regions of interest; some studies used the raw data without any form of feature selection whereas others performed a number of transformations on the data to select relevant features; and different studies used different DL architectures. Such methodological variability means that, at present, the reliability and replicability of the existing results remain unclear.

4.2. The promise of convolutional neural networks

CNNs are a particular type of feedforward neural network inspired by how the human visual cortex process information. Over the past decade, CNNs have been breaking records in computer vision across several competitions, making this approach a very promising one (Krizhevsky et al., 2012). Consistent with this, our review has shown that CNNs have generated the most encouraging results in the context of neuroimaging. In its raw form, neuroimaging data comprises millions of voxels. Considering the current computational resources available, putting all voxel intensities through a fully connected network would lead to an unfeasible number of weights to be estimated. Two intrinsic properties of CNNs – weight sharing and local connectivity – result in a significantly reduced number of weights, making it computationally possible to run the network at the voxel-level. Although in neuroimaging CNNs have only been used to examine MCI and AD patients, the accuracies of the studies published so far have been consistently high (i.e. $\geq$95% for AD and $\geq$86% for MCI versus controls). High accuracies have been observed with different modalities including structural MRI (Gupta et al., 2013; Hosseini-Asl et al., 2016; Payan and Montana, 2015), resting-state fMRI (Sarraf and Tofighi, 2016) and CT imaging (Gao and Hui, 2016), as well as with small (Gao and Hui, 2016; Sarraf and Tofighi, 2016) and large (Gupta et al., 2013; Hosseini-Asl et al., 2016; Payan and Montana, 2015) sample sizes. Hosseini-Asl et al. (2016) used an alternative and interesting approach which involved pre-training a CNN in one Alzheimer's dataset (CADDementia) and then fine-tuning and testing it in another dataset from the same diagnostic group (ADNI). The results were very promising for both 2-way and 3-way classifications (HC vs. AD; HC vs. MCI; AD vs. MCI; and HC vs. AD vs. MCI), although it should be noted that the ADNI sample was of modest size. Taken together, these results are in line with the successful performances of CNN-based models reported in other scientific areas, and highlight CNNs as a promising tool in neuroimaging.

4.3. From binary to multiclass classifications

In the context of neuroimaging, the vast majority of conventional ML studies have relied on binary classifications involving the comparison between a group of patients and a group of healthy controls (Orrù et al., 2012; Wolfers et al., 2015). This can be explained by the fact that these studies have typically employed SVM, which was originally designed for binary classification problems (Hsu and Lin, 2002). However, the real challenge for clinicians is not to differentiate between patients and controls but to develop biomarkers which could be used to choose amongst alternative diagnoses or different stages of illness progression. Looking forward, therefore, ML models will need to be able to discriminate amongst several possible alternatives in order to inform real-world clinical decision making. Many approaches have been proposed to enable SVM to handle multiclass classification problems (Fei and Liu, 2006; Hsu and Lin, 2002). However, this is still an active research area (Kumar and Gopal, 2011) and none of the proposed approaches have been tested in the context of neuroimaging. Most neuroimaging studies using SVM addressed the multiclass problem by performing several binary classifications (for example, AD vs. HC, MCI vs. HC and AD vs. MCI) or one-against-all classifications (for example, AD vs. MCI & HC and MCI vs. AD & HC). DL however, requires less technical effort to perform multiclass comparisons, and therefore could provide a solution to this issue. This is mainly due to the use of the so-called softmax function in the output layer, which can be considered an extension of the binary logistic regression to several classes. Here the output reflects the probability of belonging to each class, which is a more intuitive index of class membership than some of the most sophisticated indices being developed for SVM multiclass solutions (Fei and Liu, 2006). In light of its suitability for multiclass classification, a number of studies have used DL to carry out 3 or 4-way classifications between different disorder subtypes or different stages of illness. For example, three of these studies were able to classify children into healthy controls and three ADHD subtypes (inattentive, hyperactive and combined) (Hao et al., 2015; Kuang and He, 2014; Kuang et al., 2014). Notably, there is also preliminary evidence for the use of DL to distinguish between individuals at no imminent risk of dementia, those identified at risk who will and will not develop dementia, and those with established Alzheimer's disease (Liu et al., 2015a; Liu et al., 2014; Suk et al., 2015b). These are encouraging findings, as they highlight how DL could help bridge the existing gap between neuroimaging findings and real-world clinical practice.

4.4. Is deep learning superior to conventional machine learning?

Despite the success of DL in several scientific areas, the superiority of this analytical approach in neuroimaging is yet to be demonstrated. On the one hand, DL has been described as a potentially more powerful approach than conventional shallow ML, as it is capable of learning highly intricate and abstract patterns from the data, which can particularly useful in the case of brain-based disorders (Plis et al., 2014). On the other hand, given that neuroimaging data is very high-dimensional, the nonlinear approach of DL might not be advantageous as there are not enough data points to extract meaningful nonlinear patterns from the data, whereas the linear approach employed in conventional shallow ML might be more appropriate. Here we tried to clarify this issue by systematically examining the difference in performance between DL and conventional shallow ML in studies which used both approaches. A total of twenty-five studies reported classification accuracy for both DL and conventional shallow ML, with the latter being a kernel-based method, either SVM or MKL. For the majority of these studies DL performed better than conventional shallow ML as shown in Fig. 5, and in some cases the difference was by a reasonable margin (e.g. Han et al., 2015; Plis et al., 2014; Suk and Chen, 2013).

From the available evidence, it is not clear whether DL tends to perform better under specific circumstances, for example depending on the modality type or the sample size. However, our systematic review provides anecdotal evidence that studies combining imaging and non-imaging data tend to have a larger margin in favour of DL (see Fig. 6). This is consistent with the notion that the association between brain abnormalities and cognitive symptoms, for example, is likely to exist at a deep and abstract level, and as such can be captured more effectively by DL methods than traditional shallow ML methods (Plis et al., 2014).

We know that the application of traditional shallow ML methods to neuroimaging data leads to higher and more stable accuracies as the sample size increases (Nieuwenhuis et al., 2012). One would expect this to be especially true for DL: since a deep model is inherently more complex than conventional shallow ML models, larger

sample sizes should be needed to compensate for the greater number of parameters to be estimated and to take full advantage of DL's ability to detect highly intricate and abstract patterns in the data. We were therefore expecting to see an increase in the margin by which DL outperforms kernel-based methods as sample sizes increase. Such increase however was not observed, as the pattern of difference in performance did not seem to vary systematically with sample size; one possibility is that larger sample sizes than those used in the existing literature would be required to detect increases in the margin by which DL outperforms kernel-based methods.

In conclusion, our review suggests that, overall, DL performs better than conventional shallow ML. In light of the increasing interest in DL, however, we cannot exclude a publication bias which favoured studies showing the superiority of this new analytical approach relative to conventional shallow ML methods (Boulesteix et al., 2013). As the number of studies applying DL to neuroimaging data increases, a thorough assessment of publication bias would be useful to establish the reliability of this initial trend in favour of DL.

4.5. Interpretability of DL in neuroimaging

Despite having demonstrated state-of-the-art performances across several fields, DL has been under scrutiny for its lack of transparency during the learning and testing processes (Alain and Bengio, 2016; Lou et al., 2012; Yosinski et al., 2015). For example, deep neural networks have been referred to as a "black box" in contrast with other techniques, such as logistic regression, which are less complex and more intuitive. Such lack of transparency has important implications for the interpretability of the results when DL is applied to neuroimaging data. Due to the multiple nonlinearities, it can be challenging to trace the consecutive layers of weights back to the original brain image in order to identify which features (e.g. regions) are providing the greatest contribution to classification (Suk et al., 2015a). This information however would be useful in the context of clinical neuroimaging where the aim is not only to detect but also localise abnormalities. A first potential issue is that a model with an excellent performance may be using irrelevant features (e.g. orientation of the images, imaging artefacts), as oppose to clinically meaningful information (e.g. regional grey matter, connectivity between different brain regions), to classify participants. A second potential issue is that an accurate model which provides no information about the underlying neuroanatomical or neurofunctional alterations would be of limited clinical utility, for example with respect to treatment development and optimization.

Despite its complex inner workings which make the visualization and interpretation of the weights challenging, DL *can* be used in a way which enables transparency. This is illustrated by several neuroimaging studies included in this review that did report the most important features (e.g., Deshpande et al., 2015; Kim et al., 2016; Liu et al., 2014; Suk et al., 2016). However, these studies used a variety of approaches to isolate the most informative features, and at present there is no standard and intuitive method for visualizing weights or interpreting latent feature representations (Suk et al., 2015a). This has motivated several attempts to develop new and intuitive ways of enhancing the interpretability of DL within the recent literature (e.g., Grün et al., 2016; Samek et al., 2015; Simonyan et al., 2013; Yosinski et al., 2015; Zeiler and Fergus, 2014). There are two main methodological approaches to address this issue, including input modification methods and deconvolution methods. Input modification methods are visualization techniques that involve the systematic modification of the input and the measurement of any resulting changes in the output as well as in the activation of the artificial neurons in the intermediate layers of the network. An example of these methods is the so-called occlusion method (Zeiler and Fergus, 2013) which involves covering portions of the input image up to find the areas of the input data that influence the probability of the output classes. In contrast, deconvolution methods aim to determine the contribution of one or more features of the input data to the output. This involves selecting an activation of interest in an output neuron and then computing the contribution of each neuron in the next lower layers to this activation. Here a number of strategies are available to model the nonlinearities present across the layers, for example, deconvnet (Zeiler and Fergus, 2013) and guided backpropagation (Springenberg et al., 2014).

4.6. The challenge of overfitting

Overfitting is arguably one of the main challenges in ML. Given their inherent complexity, DL networks are particularly prone to overfitting, i.e., learning irrelevant fluctuations in the data that limit generalizability. Not surprisingly, different approaches to address this issue, known as regularization strategies, have been developed and are now present in most DL algorithms. In section 2.1.4 we described some of the most commonly used regularization strategies applied to modern DL, namely weight decays and dropout. As expected, several studies reviewed here have used some form of regularization. The majority (e.g., Hosseini-Asl et al., 2016; Kim et al., 2016; Liu et al., 2015a) have employed the L1 or L2 norms, which prevent overfitting by penalizing very low or very high weight values. At least one study (Li et al., 2014) employed dropout, where a random number of nodes and respective connections are temporarily removed to extract different sets of features that can independently produce a useful output. The importance of regularization strategies in DL could potentially account for the fact that Munsell and colleagues, who trained 4- and 5-hidden layer models (for inferring diagnostic and treatment outcome, respectively) without using any form of regularization, reported such low performance for DL (Munsell et al., 2015).

An additional approach for minimising the risk of overfitting involves reducing the dimensionality of the data before inputting them into the model. A possible way of achieving this is by extracting region- or patch-level features (as opposed to using voxel-level data). Using different types of features (whether voxel, patch or region) can have implications for how detailed the information inputted into the model is (for example, voxel-level features are very detailed, and also very noisy; region-level features on the other hand, ignore more localized patterns and are less sensitivity to noise). Another option to reduce dimensionality is feature selection. Feature selection is common in conventional ML, where linear methods such as principal component analysis, independent component analysis or elastic net, are used to select the most discriminating features that are then fed to a classifier. However, the use of conventional feature selection methods prior to a DL model seems counterintuitive, since one of the main advantages of DL is the ability to learn, through a purely data-driven method, the most useful features for classification. Several studies reported in this review have attempted to reduce the dimensionality of the data by extracting region- or patch-level features, using feature selection, or combining the two approaches. We note, however, that all CNN-based models were applied to voxel-level data without being preceded by any form of feature selection and yet reported consistently high performances on unseen data. This suggests that DL, and CNN-models and particular, can perform well with neuroimaging data without the requirement to downsize or even preprocess the data. For example, Hosseini-Asl et al. (2016) achieved high levels of accuracy after applying a CNN to voxel-level data without any preprocessing or even skull stripping of the images. This finding has potential implications for the development of clinical tools, as it suggests that it might be possible to apply DL to raw neuroimaging data, thereby saving time as well as technical resources.

4.7. Technical expertise and computational requirements

The studies reviewed in this article employed a wide range of DL architectures and hyperparameters. Such flexibility is what makes DL a very powerful tool but comes at a potentially high cost. The number of layers, the number of nodes within each layer and the activation function of each node are only a few examples of a long list of variables one has to consider when designing and optimizing a DL model. Automated optimization strategies are not yet widely available, making optimisation a manual process that requires a great deal of technical expertise and is potentially prone to subjective bias. Since the number of parameters to be estimated is very large, the computational requirements of DL are also more demanding than those of conventional ML methods. For example, Kim et al. (2016) reported that the estimation of a DL model with three hidden layers took 100 times longer than the estimation of a standard SVM model (~3.3 days vs. 0.8 h). However, with the fast-growing availability of graphical processing units (GPUs), the application of DL to neuroimaging data is likely to become less and less time-consuming in the future.

5. Conclusions and future directions

While still in its initial stages, the application of DL in neuroimaging has shown promising results and has the potential of leading to fundamental advances in the search for imaging-based biomarkers of psychiatric and neurologic disorders. Nevertheless, several improvements will be required before the full potential of DL in neuroimaging can be achieved. Firstly, given the complexity of DL models, we need to move away from studies with small to modest sample sizes in favour of much larger cohorts. A possible way of achieving this is through multi-centre collaborations, in which data is collected using the same recruitment criteria and scanning protocols across sites. A further way of increasing the sample size is through multi-site data sharing initiatives, such as ADNI for Alzheimer's disease and ADHD-200 for ADHD. Secondly, the integration of CNN and recurrent neural networks (i.e. networks that allow the processing of data with sequential inputs such as videos or speech) is likely to lead to significant advances in DL in the next few years (Donahue et al., 2015). In neuroimaging, this integration could be particularly useful for analysing fMRI data, as it would allow the detection of intricate spatial patterns while simultaneously modelling the temporal component of the BOLD signal. Thirdly, we anticipate that an increasing number of neuroimaging studies will make use of transfer learning, which involves using previously learned features from a large sample of similar enough images. This could help tackle the curse of dimensionality – a common problem in neuroimaging studies of brain disorders (Gupta et al., 2013; Hosseini-Asl et al., 2016). Evidence from vision science, where deeper models such as VGG net (Simonyan and Zisserman, 2014), residuals networks (He et al., 2015) and Inception-v4 (Szegedy et al., 2016) are achieving the highest performances, suggests that transfer learning could be particularly useful when deeper models are employed. Fourthly, we suggest that the so-called augmentation technique – which it is commonly used in computer vision – could be useful in the context of neuroimaging. This technique involves increasing the sample size by applying transformations to the data (e.g., rotation, shear, scaling), and then train a model that is invariant to such transformations. The use of augmentation could also address the issue of modest sample sizes and lead to a decrease in prepossessing time (because steps such as rotation may become redundant). Finally, the use of DL to predict continuous scores is another interesting area for further research with potential clinical applicability, following the encouraging results obtained using conventional ML methods (e.g. Gong et al., 2014; Stonnington et al., 2010; Tognin et al., 2014). So far, only one study has used DL to predict clinical scores from structural MRI scans in patients with Alzheimer's disease (Brosch and Tam, 2013).

In conclusion, the capacity of DL models to learn complex and abstract representations through nonlinear transformations, makes this a promising approach to single subject prediction in neuroimaging. While there are still important challenges to overcome, the findings reviewed here provide preliminary evidence supporting the potential role of DL in the future development of diagnostic and prognostic biomarkers of psychiatric and neurologic disorders.

Acknowledgements

Sandra Vieira is supported by a PhD studentship from the Fundação para a Ciência e a Tecnologia (FCT), research grant SFRH/BD/103907/2014. Walter H.L. Pinaya gratefully acknowledges support from FAPESP (Brazil), grant #2013/05168-7, São Paulo Research Foundation. Andrea Mechelli is supported by the Medical Research Council (ID99859).

References

Alain, G., Bengio, Y., 2016. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644.
Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V., Diener-West, M., 2004. The use of overall accuracy to evaluate the validity of screening or diagnostic tests. J. Gen. Intern. Med. 19, 460–465.
Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2016. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage, 137–165.
Bengio, Y., 2009. Learning deep architectures for AI. Found. Trends® Mach. Learn. 2, 1–127.
Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. Adv. Neural Inf. Process. Syst., 2546–2554.
Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kotter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc. Natl. Acad. Sci. 107, 4734–4739.
Boulesteix, A.L., Lauer, S., Eugster, M.J., 2013. A plea for neutral comparison studies in computational sciences. PLoS One 8, e61562.
Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. Proceedings of the IEEE 20th International Conference on Pattern Recognition, 3121–3124.
Brosch T., Tam R., Alzheimer's Disease Neuroimaging Initiative, 2013. Manifold learning of brain MRIs by deep learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 633–640. Springer Berlin Heidelberg.
Cabral, C., Kambeitz-Ilankovic, L., Kambeitz, J., Calhoun, V.D., Dwyer, D.B., von Saldern, S., Urquijo, M.F., Falkai, P., Koutsouleris, N., 2016. Classifying schizophrenia using multimodal multivariate pattern recognition analysis: evaluating the impact of individual clinical profiles on the neurodiagnostic performance. Schizophr. Bull. 42, S110–S117.
Calhoun, V.D., Sui, J., 2016. Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness. Biol. Psychiatry: Cogn. Neurosci.Neuroimag. 1, 230–244.
Chen, Y., Shi, B., Smith, C.D., Liu, J., 2015. Nonlinear Feature Transformation and Deep Fusion for Alzheimer's Disease Staging Analysis. In: International Workshop on Machine Learning in Medical Imaging, 304–312. Springer International Publishing.
Deshpande, G., Wang, P., Rangaprakash, D., Wilamowski, B., 2015. Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. IEEE Trans. Cybernet. 45, 2668–2679.
Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2625–2634.
Fei, B., Liu, J., 2006. Binary tree of SVM: a new fast multiclass training and classification algorithm. IEEE Trans. Neural Netw. 17, 696–704.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U. S. A. 102, 9673–9678.
Gao, X.W., Hui, R., 2016. A deep learning based approach to classification of CT brain images. In: Science and Information Conference, London, UK.
Gelbart, M.A., Snoek, J., Adams, R.P., 2014. Bayesian optimization with unknown constraints. arXiv preprint arXiv:1403.5607.
Gong, Q., Li, L., Du, M., Pettersson-Yeo, W., Crossley, N., Yang, X., Li, J., Huang, X., Mechelli, A., 2014. Quantitative prediction of individual psychopathology in trauma survivors using resting-state FMRI. Neuropsychopharmacology 39, 681–687.
Grün, F., Rupprecht, C., Navab, N., Tombari, F., 2016. A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks. arXiv preprint arXiv:1606.07757.
Gupta, A., Ayhan, M., Maida, A., 2013. Natural image bases to represent neuroimaging data. International Conference on Machine Learning, 987–994.
Han X., Zhong Y., He L., Philip S.Y., Zhang L., 2015. The unsupervised hierarchical convolutional sparse auto-encoder for neuroimaging data classification. In: International Conference on Brain Informatics and Health, 156–166. Springer International Publishing.
Hao, A.J., He, B.L., Yin, C.H., 2015. Discrimination of ADHD children based on deep bayesian network. 2015 International Conference on Biomedical Image and Signal Processing, 1–6.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York, NY.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.
Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.
Hosseini-Asl, E., Gimel'farb, G., El-Baz, A., 2016. Alzheimer's Disease Diagnostics by a Deeply Supervised Adaptable 3D Convolutional Network. arXiv preprint arXiv:1607.00556.
Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13, 415–425.
Hu, C., Ju, R., Shen, Y., Zhou, P., Li, Q., 2016. Clinical decision support for Alzheimer's disease based on deep learning and brain network. Proceedings of the IEEE International Conference on Communications, 1–6.
Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Della Penna, S., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., Handwerker, D.A., Keilholz, S., Kiviniemi, V., Leopold, D.A., de Pasquale, F., Sporns, O., Walter, M., Chang, C., 2013. Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage 80, 360–378.
Kennedy, D.P., Courchesne, E., 2008. The intrinsic functional organization of the brain is altered in autism. Neuroimage 39, 1877–1885.
Kim, J., Calhoun, V.D., Shim, E., Lee, J.H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage 124, 127–146.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems., pp. 1097–1105.
Kuang, D., He, L., 2014. Classification on ADHD with deep learning. Proceedings of the International Conference on Cloud Computing and Big Data, 27–32.
Kuang, D., Guo, X., An, X., Zhao, Y., He, L., 2014. Discrimination of ADHD based on fMRI data with deep belief network. International Conference on Intelligent Computing, 225–232.
Kumar, M.A., Gopal, M., 2011. Reduced one-against-all method for multiclass SVM classification. Expert Syst. Appl. 38, 14238–14248.
Larochelle, H., Erhan, D., Courville, A., Bergstra, J., Bengio, Y., 2007. An empirical evaluation of deep architectures on problems with many factors of variation. Proceedings of the 24th International Conference on Machine Learning, 473–480.
Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., Ng, A., 2012. Building high-level features using large scale unsupervised learning. International Conference on Machine Learning 103.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Li, F., Tran, L., Thung, K.H., Ji, S., Shen, D., Li, J., 2014. Robust deep learning for improved classification of AD/MCI patients. International Workshop on Machine Learning in Medical Imaging, 240–247.
Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014. Early diagnosis of Alzheimer's Disease with deep learning. IEEE 11th International Symposium on Biomedical Imaging, 1015–1018.
Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., 2015a. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans. Biomed. Eng. 62, 1132–1140.
Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D.D., 2015b. Multi-phase feature representation learning for neurodegenerative disease diagnosis. Australasian Conference on Artificial Life and Computational Intelligence, 350–359.
McCulloch, W., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 7, 115–133.
Mechelli, A., Prata, D., Kefford, C., Kapur, S., 2015. Predicting clinical response in people at ultra-high risk of psychosis: a systematic and quantitative review. Drug Discovery Today 20, 924–927.
Milham, M.P., Fair, D., Mennes, M., Mostofsky, S.H., 2012. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Front. Syst. Neurosci. 6, 62.
Moody, J., Hanson, S., Krogh, A., Hertz, J.A., 1995. A simple weight decay can improve generalization. Adv. Neural Inf. Process. Syst. 4, 950–957.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Alzheimer's disease neuroimaging initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.
Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005a. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's Dementia 1, 55–66.
Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005b. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15, 869–877.
Mulders, P.C., van Eijndhoven, P.F., Schene, A.H., Beckmann, C.F., Tendolkar, I., 2015. Resting-state functional connectivity in major depressive disorder: a review. Neurosci. Biobehav. Rev. 56, 330–344.
Munsell, B.C., Wee, C.Y., Keller, S.S., Weber, B., Elger, C., da Silva, L.A.T., Nesland, T., Styner, M., Shen, D., Bonilha, L., 2015. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. Neuroimage 118, 219–230.
Nieuwenhuis, M., van Haren, N.E., Pol, H.E.H., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. Neuroimage 61, 606–612.
Nowlan, S.J., Hinton, G.E., 1992. Simplifying neural networks by soft weight-sharing. Neural Comput. 4, 473–493.
Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36, 1140–1152.
Page, A., Turner, J.T., Mohsenin, T., Oates, T., 2014. Comparing raw data and feature extraction for seizure detection with deep learning methods. International Florida Artificial Intelligence Research Society Conference.
Payan, A., Montana, G., 2015. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv preprint arXiv: 1502.02506.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.
Pettersson-Yeo, W., Benetti, S., Marquand, A.F., Dell'Acqua, F., Williams, S.C.R., Allen, P., Prata, D., McGuire, P., Mechelli, A., 2013. Using genetic: cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. Psychol. Med. 43, 2547–2562.
Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8, 1–11.
Radua, J., Borgwardt, S., Crescini, A., Mataix-Cols, D., Meyer-Lindenberg, A., McGuire, P.K., Fusar-Poli, P., 2012. Multimodal meta-analysis of structural and functional brain changes in first episode psychosis and the effects of antipsychotic medication. Neurosci. Biobehav. Rev. 36, 2325–2333.
Samek, W., Binder, A., Montavon, G., Bach, S., Müller, K.R., 2015. Evaluating the visualization of what a deep neural network has learned. arXiv preprint arXiv:1509.06321.
Sarraf, S., Tofighi, G., 2016. Classification of Alzheimer's Disease using fMRI Data and Deep Learning Convolutional Neural Networks. arXiv preprint arXiv:1603.08631.
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117.
Schultz, C.C., Fusar-Poli, P., Wagner, G., Koch, K., Schachtzabel, C., Gruber, O., Sauer, H., Schlösser, R.G., 2012. Multimodal functional and structural imaging investigations in psychosis research. Eur. Arch. Psychiatry Clin. Neurosci. 262, 97–106.
Sheffield, J.M., Barch, D.M., 2016. Cognition and resting-state functional connectivity in schizophrenia. Neurosci. Biobehav. Rev. 61, 108–120.
Simonyan, K., Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.
Springenberg, J.T., Dosovitskiy, A., Brox, T., and Riedmiller, M., 2014. Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806.
Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Stonnington, C.M., Chu, C., Klöppel, S., Jack, C.R., Ashburner, J., Frackowiak, R.S., 2010. Alzheimer Disease Neuroimaging Initiative. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51, 1405–1413.
Suk, H.I., Shen, D., 2013. Deep learning-based feature representation for AD/MCI classification. International Conference on Medical Image Computing and Computer-Assisted Intervention, 583–590.
Suk, H.I., Lee, S.W., Shen, D., 2014. Alzheimer's Disease Neuroimaging Initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 101, 569–582.

Suk, H.I., Lee, S.W., Shen, D., 2015a. Alzheimer's disease neuroimaging initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859.
Suk, H.I., Lee, S.W., Shen, D., 2015b. Alzheimer's Disease Neuroimaging Initiative. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. Brain Struct. Funct., 1–19.
Suk, H.I., Wee, C.Y., Lee, S.W., Shen, D., 2016. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. Neuroimage 129, 292–307.
Szegedy, C., Ioffe, S., Vanhoucke, V., 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261.
Tognin, S., Pettersson-Yeo, W., Valli, I., Hutton, C., Woolley, J., Allen, P., McGuire, P., Mechelli, A., 2014. Using structural neuroimaging to make quantitative predictions of symptom progression in individuals at ultra-high risk for psychosis. Front. Psychiatry 4, 187.
Valli, I., Marquand, A.F., Mechelli, A., Raffin, M., Allen, P., Seal, M.L., McGuire, P., 2016. Identifying individuals at high risk of psychosis: predictive utility of Support Vector Machine using structural and functional MRI data. Front. Psychiatry 7.
van der Meer, L., Costafreda, S., Aleman, A., David, A.S., 2010. Self-reflection and the brain: a theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. Neurosci. Biobehav. Rev. 34 (6), 935–946.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408.
Willette, A.A., Calhoun, V.D., Egan, J.M., Kapogiannis, D., 2014. Alzheimer' s Disease Neuroimaging Initiative. Prognostic classification of mild cognitive impairment and Alzheimer's disease: MRI independent component analysis. Psychiatry Res.: Neuroimag. 224, 81–88.
Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci. Biobehav. Rev. 57, 328–349.
Yang, Z., Zhong, S., Carass, A., Ying, S.H., Prince, J.L., 2014. Deep learning for cerebellar ataxia classification and functional score regression. International Workshop on Machine Learning in Medical Imaging, 68–76.
Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H., 2015. Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.
Yung, A.R., Yuen, H.P., McGorry, P.D., Phillips, L.J., Kelly, D., Dell'Olio, M., Francey, S.M., Cosgrave, E.M., Killackey, E., Stanford, C., Godfrey, K., Buckby, J., 2005. Mapping the onset of psychosis: the comprehensive assessment of at-risk mental states. Aust. N. Z. J. Psychiatry 39, 964–971.
Zarogianni, E., Moorhead, T.W., Lawrie, S.M., 2013. Towards the identification of imaging biomarkers in schizophrenia: using multivariate pattern classification at a single-subject level. NeuroImage: Clin. 3, 279–289.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In European Conference on Computer Vision, 818–833. Springer International Publishing.
Zhang, D., Shen, D., 2012. Alzheimer's Disease Neuroimaging Initiative. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7, e33182.